

# Universality splitting in distribution of number of miRNA co-targets

Mahashweta Basu<sup>1</sup>, Nitai P. Bhattacharyya<sup>2</sup>, P. K. Mohanty<sup>1</sup>

<sup>1</sup>Condensed Matter Physics Division, Saha Institute of Nuclear Physics, 1/AF Bidhan Nagar, Kolkata 700064, India.

<sup>2</sup>Crystallography and Molecular Biology Division,  
Saha Institute of Nuclear Physics, 1/AF Bidhan Nagar, Kolkata 700064, India.

In a recent work [arXiv:1307.1382] it was pointed out that the link-weight distribution of microRNA (miRNA) co-target network of a wide class of species are universal up to scaling. The number cell types, widely accepted as a measure of complexity, turns out to be proportional to these scale-factor. In this article we discuss additional universal features of these networks and show that, this universality splits if one considers distribution of number of common targets of three or more number of miRNAs. These distributions for different species can be collapsed onto two distinct set of universal functions, revealing the fact that the species which appeared in early evolution have different complexity measure compared to those appeared late.

MicroRNAs are small non-coding single stranded RNAs of about 22 nucleotides long and act as a secondary regulator of gene expression [1–3]. They are transcribed from the DNA from either inter or intra genomic region [4, 5] and bind to the UTRs of some of the mRNAs to inhibit their functionality [6]. In effect, the respective proteins are produced less compared to the situation when miRNAs are absent [5, 7]. Even being secondary regulators (transcription factors being the primary ones) miRNAs are seen as potential therapeutic targets for treatment of cancer [8, 9] and other disease [10, 11]. There are large number of databases [12, 13] which predict short genomic sequences which might be acting as a miRNAs. Substantial effort [14–19] has also been given in predicting the targets of these miRNAs.

The miRBase database [12] predicts that there are about 851 miRNAs for *Homo sapiens*. A web resource MicroCosm Targets [19] provides computationally predicted targets of microRNAs across many species (for example human miRNAs target about 950 mRNAs out of total 35864). Experimental validation of these predictions are, however, largely lacking. It is believed [20] that specific biological functions and processes are possibly carried out by groups of miRNAs, through noise reduction, than individual ones. It is thus important to look for combinatorial regulation [21]. Recent studies on miRNA [22, 23] co-target network for *Homo sapiens* reveal the miRNA groups and obtained the respective functions. The miRNA co-target network is formed by joining a pair of miRNA by link and associating the number of co-targets as the weight of the link. Community structures of these densely weighted networks are then obtained using certain modularization algorithms [22, 24].

Surprisingly the weight distribution of these networks show amazing universal features, which extends over many species classes, families and genera [25]. In other words the distribution function  $P(w)$  of the number of co-targets  $w$  is found to be a scaled form of an universal function; the species are characterized by a unique scale factor  $\lambda$ . It was also observed that the scale factor is proportional to the number of cell types of the respective

species, and thus, can be considered as a measure of the complexity. A simple random-target model, where miRNAs of a species target a fixed number of mRNAs, could produce the universal function reasonably well. Thus, animal specificity is not resolved through pair-wise co-targets and from these networks one expects to obtain generic functions common to species of wide range.

In this article, we propose that, if one consider number of common targets of three or more miRNAs of a given species, the distribution function show *two* distinct groups of animals. We also observe that this sub-groups are consistent with the natural partition of the species into two groups obtained from the bi-model distribution of number of miRNAs.

For completeness, first we discuss how to obtain the miRNA co-target distributions for different species. We consider all the species whose miRNA targets have been predicted by MicroCosm Targets [19]. The complete list of species, along with the number of miRNAs  $M$  and mRNAs  $N$  are given in Table I. Let us denote the miRNAs of a species as  $\{m_1, m_2 \dots m_M\}$ . To construct the co-target of miRNA multiplets of size  $k$ , we first take  $k$  distinct miRNAs  $\{m_{i_1}, m_{i_2}, \dots m_{i_k}\}$  from the set of  $M$ . Thus in total, there are  $C_k^M$  multiplets. Then from the target database, we find the targets which are common to the first two miRNAs  $m_{i_1}$  and  $m_{i_2}$  and denote the number of common targets as  $w_2$ . These  $w_2$  targets are then compared with the targets of next miRNA  $m_{i_3}$  to find  $w_3$  which is now the number of common targets of the triplet  $\{m_{i_1}, m_{i_2}, m_{i_3}\}$ . This process is iterated until one obtains  $w_k$ . The nonzero  $w_k$ s obtained through this process are now considered for obtaining the distribution function  $P_k(w) \equiv P(w_k)$ . Clearly, for  $k = 2$ , the distribution  $P_2(w)$  is equivalent to  $P(w)$ , the link-weight distribution of miRNA co-target networks discussed in Ref. [25]. In Fig. 1(a) we have shown  $P_2(w)$  for four different species, namely *Homo sapiens* (Human), *Bos taurus* (bovine), *Xenopus tropicalis* (western clawed frog) and *C. elegans*. Note that, the number of common genes targeted by any miRNA pair does not depend on the total number miRNAs of the species, however depends

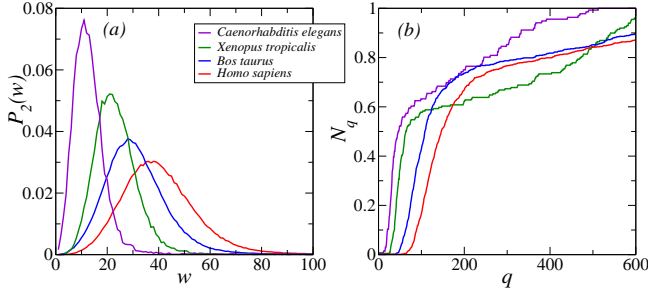


FIG. 1. (a) Link-weight distribution  $P_2(w)$  and (b) number of components  $N_q$  are shown for four different species ( *C. elegans*, *Xenopus tropicalis*, *Bos taurus* and *Homo sapiens*).

on the number of genes  $N$ . To understand the shift in peak position and the change in width (or variance) of  $P_2(w)$ , a simple model has been proposed in Ref. [25]. It was shown that the peak position is proportional to square of the average number of genes targeted by the miRNAs and inversely proportional to  $N$ . Interestingly,  $P_2(w)$ s for different species could be collapsed onto each other by aligning their peaks at origin and rescaling the  $x$ - and  $y$ - axis suitably by a single factor  $\lambda$  which is species dependent. It was observed that the scale factors are proportional to the number of cell-types and serve as an equivalent measure of complexity of the species.

In the following we consider certain other universal features of the miRNA co-target network. Mookherjee *et. al.* [22] have constructed the miRNA co-target network of *Homo sapiens* from a adjacency matrix  $W$  with elements  $w_{ij}$  same as the number of co-targets of miRNA pair  $i$  and  $j$ . The human miRNA co-target network was found to be fully connected with large variation in link-weights - some as large as 1282 and as small as 1. They argued that links with small weights are rather un-important and the network can be made simple by erasing links whose weights are smaller than a pre-specified value  $q$ . In this case, the network breaks into  $N_q$  number of disconnected components;  $N_q$  being a non-decreasing function of  $q$  with  $N_0 = 1$  (as all miRNAs are connected at  $q = 0$ ). The variation of  $N_q$  with  $q$  for *Homo sapiens* and other three species are shown in Fig. 1(b). Let us define density of components

$$\nu_q = \frac{N_q - 1}{M - 1}. \quad (1)$$

Evidently, as  $q$  is increased,  $\nu_q$  picks up a non-zero value at some  $q = q_c$  (when the network starts breaking up). Mookherjee *et. al.* [22] have claimed that the optimum network, that does not contain irrelevant links nor loses the network functionality, occurs at a value of  $q^* = 103$  where the breaking rate  $\frac{d\nu_q}{dq}$  is maximum. The largest component at  $q = q^*$ , which contains 429 miRNAs, provide all essential regulations. This group of miRNAs consists of several small clusters which are found to be tissue, pathway, diseases specific.

We revisit co-target networks for human and 21 other species and find that the density of components  $N_q$  also shows certain other universal features. Firstly, all the networks are found to be fully connected, with unit clustering coefficient and diameter. Further,  $\nu_q$  for all the species show data collapse, *i.e.* one can write  $\nu_q = \mathcal{F}(A(q - q_c))$  where  $q_c$  is the critical threshold where the network starts breaking into dis-joint components and  $A$  is a scale factor. We find that  $\nu_q$  is scale free near  $q = q_c$ ,

$$\nu_q \sim (q - q_c)^\beta. \quad (2)$$

In Fig. 2 we have plotted  $\nu_q$  as a function of  $(q - q_c)$  in log scale where (a) corresponds to species with less than 250 miRNAs and (b) corresponds to the rest. The  $x$ -axis is rescaled here to obtained the collapse. In fact figures (a) and (b) could be collapsed onto each other, but they are shown as separate figures only emphasize that the data for species with small number of miRNAs are comparatively noisy. Our best estimate is  $\beta = 2$ ; a straight line with slope  $\beta = 2$  is drawn in both figures for comparison. Interestingly, we find that critical threshold  $q_c$  is also proportional to the number of cell-types (see inset of Fig 2(b)) and thus, it can also be considered as an equivalent measure of complexity.

Upto this point, we have discussed that the network formed by the common targets of miRNA-pairs are universal in many ways: the link weight distribution  $P(w)$  and the density of components  $\nu_q$  near the breakdown point  $q = q_c$  for different species across a wide class are only scaled forms of respective universal functions. It is rather surprising that the species specificity show up as a scale factors. One thus expects that biological functions co-regulated by miRNA pairs are possibly less specific and occurs widely across many species. To reveal more specific functions, which might be strongly species dependent, we try to find out common targets of more number of miRNAs, by taking  $k > 2$ . Note that for  $k > 2$  the

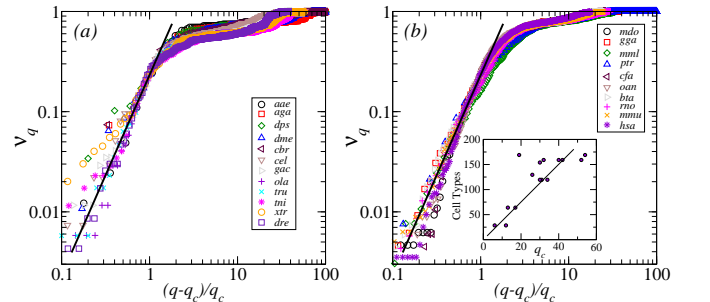


FIG. 2.  $\nu_q$  vs  $(q - q_c)$  for (a) Group-I and (b) Group-II, when scaled suitably, show data collapse. The scaling function is scale free near  $q = q_c$  with exponent  $\beta = 2$  (solid line). Inset of (b) shows that the  $q_c$  is proportional to the number of cell types of the respective species; the proportionality constant, from the best fitted line, is 3.76.

TABLE I. List of species and corresponding parameters.

Species (short name) Group-I	M	N	MYA	$q_c$	Species (short name) Group-II	M	N	MYA	$q_c$
<i>Aedes aegypti</i> (aee)	82	16059	285	14	<i>Monodelphis domestica</i> (mdo)	644	26013	-	26
<i>Anopheles gambiae</i> (aga)	82	12708	-	13	<i>Gallus gallus</i> (gga)	651	20842	55	30
<i>Drosophila pseudoobscura</i> (dps)	88	12416	-	14	<i>Macaca mulatta</i> (mml)	656	32302	6.5	57
<i>Drosophila melanogaster</i> (dme)	93	15416	375	17	<i>Pan troglodytes</i> (ptr)	662	29355	2.7	19
<i>Caenorhabditis briggsae</i> (cbr)	135	13785	-	6	<i>Canis familiaris</i> (cfa)	668	23628	5	32
<i>Caenorhabditis elegans</i> (cel)	136	24728	415	12	<i>Ornithorhynchus anatinus</i> (ana)	668	23097	115	21
<i>Gasterosteus aculeatus</i> (gac)	172	26423	-	31	<i>Bos taurus</i> (bta)	676	25759	15	40
<i>Oryzias latipes</i> (ola)	172	23514	-	25	<i>Rattus norvegicus</i> (rno)	698	30421	55	42
<i>Takifugu rubripes</i> (tru)	173	21972	-	31	<i>Mus musculus</i> (mmu)	793	30484	55	52
<i>Tetraodon nigroviridis</i> (tni)	174	28005	420	34	<i>Homo sapiens</i> (hsa)	851	35864	0.2	54
<i>Xenopus tropicalis</i> (xtr)	199	24272	360	26					
<i>Danio rerio</i> (dre)	233	28744	420	30					

M : No. of miRNA , N : No. of target mRNAs, MYA: Million years ago (appeared),  $q_c$  : Critical threshold.

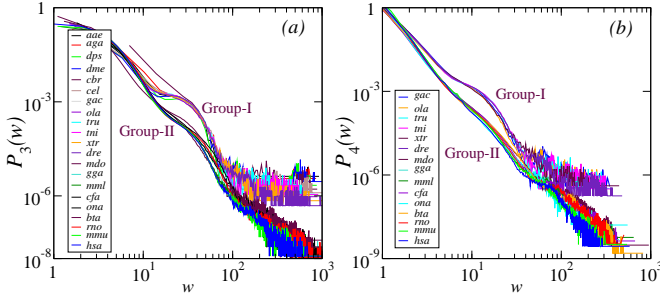


FIG. 3. (a) Distribution functions (a)  $P_3(w)$  and (b)  $P_4(w)$  of number of miRNA co-targets for different species. In (b), first 6 species of group-I are not shown as the data is noisy (for  $M$  being small). The  $y$ -axis in both plots are scaled here for obtaining the data-collapse.

number of common targets can not be simply interpreted as the the weights of some network (as they carry three or more index referring generically to tensors which, unlike matrices, does not have a network representation). In the following we study the distribution of number of common targets for  $k > 2$  number of miRNAs.

First  $k = 3$ . For each species there are  $C_3^M$  triples and the number of common targets of any three different miRNAs  $i, j > i$  and  $k > j$  is denoted as  $w_{i,j,k}$ . We find out these numbers from the miRNA target database [19], using a in house code and obtain the distribution of these numbers, denoted as  $P_3(w)$ . The same way one can obtain the distribution  $P_4(w)$  of number of common targets of  $k = 4$  miRNAs. These distribution functions are shown in Fig. 3(a) and (b) respectively. In these plots only the  $y$  axis is scaled for obtaining a data collapse. However all the data could not be collapsed on a single function; rather they split into two different scaling functions. The splitting is clearly visible for  $P_3(w)$ . For  $P_4(w)$  we find that the data is too noisy for species with small number of miRNAs (first 6 species of group-I, in table I) and it was not clear whether they just represent noise due to small number of miRNAs and targets, or there are further sub-classes (splitting). We have not shown  $P_4(w)$

for these species as they obstruct visibility of the other two collapsed-curves. Note that the plot also does not contain the distribution functions for *Pan troglodytes* as its number of targets is much lower compare to other species having nearly same number of miRNAs.

Why do we see universality splitting for mRNAs which are targeted by larger number of miRNAs ? Those mRNAs which can be regulated by more number of miRNAs take part in larger number of biological functions or pathways providing possibility of more complex gene-regulation. In this regard, it is quite possible that universality splitting reflects these complexity. The complexity of species belonging to one specific scaling function could be strikingly different from those belonging to the other scaling function. To verify, if this is indeed the case, in the following, we try to find out other measures which show the same kind of division.

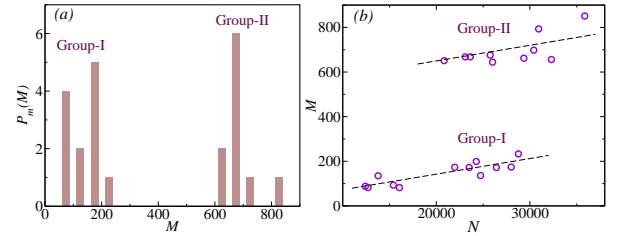


FIG. 4. (a) Distribution of number of miRNAs  $P_m(M)$ . (b) The numbers of miRNAs ( $M$ ) as a function of the number of targets ( $N$ ). Two distinct peaks in (a) and two different linear relations in (b), with same slope 0.007 (within error limits) but different intercepts 2.3 and 509.8 obtained from the best fitted line, are clear indications that the species under investigation form two different groups with respect to the number of miRNAs they have.

First we look at the distribution of number of miRNAs  $P_m(M)$ . Since, there are only a few species, we represent the distribution by a histogram by counting the number of species having  $M - 25$  to  $M + 25$  species for  $M = 25, 75, \dots, 825$  (refer to Fig. 4(a)). Clearly they show *two* distinct peaks centered about 175 and 700. It is

rather natural that less complex species, those who originate in early evolution (see Table I), have smaller number of miRNAs compared to those which are more complex. But the reason for the bi-modal structure in  $P_m(N)$  is not clear. Again, existence of these two groups can also be seen from the plot of total number of miRNAs  $M$  versus  $N$  in Fig. 4(b). Here  $M$  varies linearly with total number of target mRNAs  $N$  with a slope 0.007; however the  $y$ -intercept for group-I (2.3) is different from that of group-II (509.8). The natural division of these groups are consistent with the group of species belonging to two different scaling functions. In other words, one may say that the universality splitting is an indication that there are intricate regulatory mechanisms associated with more complex species.

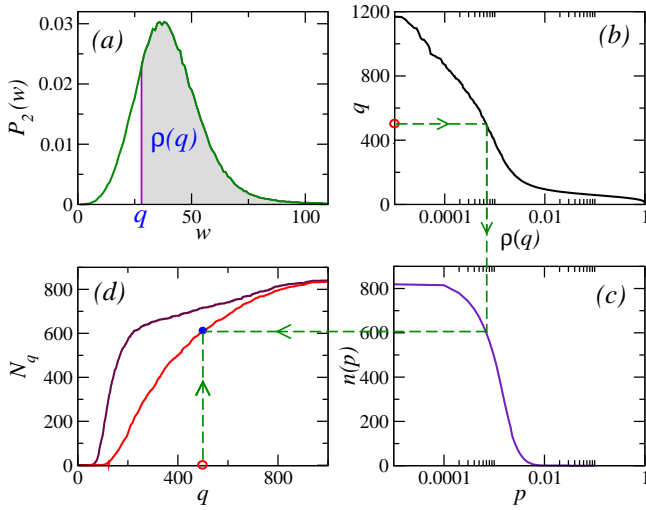


FIG. 5. (a) Link-weight distribution  $P_2(w)$  for *Homo sapiens* ( $M = 851$ ) integrated for  $w > q$  and the resulting  $\rho(q)$  is plotted in (b). (c) Number of disjoint components  $n(p)$  for a random network (851 nodes and connection probability  $p$ ). (d)  $n(\rho(q))$  for the random network is compared with  $N_q$ . Dashed-lines here describes how to obtain  $n(\rho(q))$  for a given  $q$  (see text for details).

It is thus natural to ask whether these complexity structures are also hidden somewhere in the co-target network of miRNA *pairs*. In the following we show that, even though  $\nu_q$  show universal features, it does not capture all the underlying correlations of the network. This can be done by comparing the number of components of a species with an equivalent random graph. First note that  $N_q$  and  $P_2(w)$  are related and  $N_q$  can be calculated from  $P_2(w)$ . Let

$$\rho(q) = \int_q^\infty P_2(w)dw \quad (3)$$

be the cumulative distribution of  $P_2(w)$ , which is same as the probability that the link-weight is larger than  $q$ . Since the number of components  $N_q$  is obtained by erasing all

the links with weight  $w_{ij} < q$  and assigning unit weight to all other links ( $w_{ij} > q$ ), the effective network has link density  $\rho(q)$ . We construct a random network of  $N$  nodes where the pairs are connected with probability  $\rho(q)$  and compare the number of disjoint components of this network with  $N_q$ . Let the average number of components of a random network with connection probability  $p$  and number of nodes  $N$  nodes be  $n(p)$ . Then, if the miRNA co-target networks were uncorrelated, it is expected that

$$N_q = n(\rho(q)). \quad (4)$$

In Fig. 5(d) we plot  $N_q$  for *Homo sapiens* ( $N = 851$ ) along with  $n(\rho(q))$ . The construction procedure is demonstrated in this figure. The  $P_2(w)$  for *Homo sapiens* (Fig. 5(a)) is integrated for  $w > q$  (shaded region) to obtain  $\rho(q)$  (Fig. 5(b)). Figure 5(c) shows  $n(p)$  for a random network of  $N = 851$  nodes. To obtain  $N_q$ , one starts from a given  $q$  (shown as an open circle) in Fig. 5(b), obtain  $\rho(q)$  and read out the number of components from (c) following the dashed line, and get the data point  $(q, n(\rho(q)))$  which is shown as a solid circle in Fig. 5(d). Repeating this for different values of  $q$  we obtain the expected  $N_q$  (red-line in Fig. 5(d)) for a random network. Clearly, this curve is substantially different from the actual  $N_q$  versus  $q$  curve for *Homo sapiens*, indicating presence of correlation in miRNA co-target network. Thus the co-target network is not just another random network with a specific link-weight distribution  $P_2(w)$ . These hidden correlations are uncovered in a way when one considers co-targets of three or more miRNAs.

In conclusion we have studied the distribution of number of co-targets of  $k$  number of miRNAs. For  $k = 2$  these numbers can be interpreted as the link-weight distribution  $P_2(w)$  of miRNA co-target network, which is known to have universal features. In this case,  $P_2(w)$ s for different species are only a scaled form of an universal scaling function, and the scale-factor is a measure of complexity. We show that when links of small weights (less than  $q$ ) are erased these networks breaks into several components. At the breakdown point we find an additional universal feature; the number of components show a scale free behaviour  $N_q \sim (q - q_c)^2$  and could be collapsed onto each other by rescaling of only  $x$ -axis. For  $k > 2$ , the number of co-targets does not have an graphical representation, but their distribution could also be collapsed. Surprisingly,  $P_k(w)$  for  $k > 2$  studied for 22 species show universality splitting, *i.e.*  $P_k(w)$  for a group of species collapse onto one scaling function where as the others belong to a different scaling function. The universality spiting is consistent with the bi-modal distribution of number of miRNAs and the linear dependence of number of miRNAs on the number of mRNAs (two groups have different  $y$ -intercept). The two different scaling functions are thus associated with two different class of animal, the early ones which have less number of miRNAs and less complex and the late ones which are more



complex. It remains to study, if complex regulation occurring due to the fact that if mRNAs, which are targeted by more number of miRNAs contribute to the possibility of complex regulation and new biological functions.

- 
- [1] J. Liu, Curr. Opin. Cell Biol. **20**, 214 (2008).
  - [2] D. P. Bartel, Cell **136**, 215 (2009).
  - [3] V. Ambros, Nature **431**, 350 (2004).
  - [4] D. P. Bartel, Cell **116**, 281 (2004).
  - [5] microRNA: Biology, Function and Expression, N. J. Clarke, P. Sanseau, Dna Press, 1 edition (May 2, 2006).
  - [6] K. K. Farh *et al.*, Science **310**, 1817 (2005).
  - [7] H. Dong *et al.*, Chem. Rev. **113**, 6207 (2013).
  - [8] V. Wang, W. Wu, BioDrugs. **23**, 15 (2009).
  - [9] Y. Wu *et al.*, Mol Ther Nucleic Acids. **2**, e84 (2013).
  - [10] K. F. Hansen, K. Obrietan, Neuropsychiatr Dis Treat. **9**, 1011 (2013).
  - [11] S. Hu *et al.*, Circulation. **122**, S124 (2010).
  - [12] miRBase database, [www.mirbase.org](http://www.mirbase.org)
  - [13] K. C. Miranda *et al.*, Cell **126**, 1203 (2006).
  - [14] D. P. Bartel, Cell **136**, 215 (2009).
  - [15] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, Cell **115**, 787 (2003).
  - [16] miRDB database, [www.mirdb.org/miRDB/](http://www.mirdb.org/miRDB/)
  - [17] TargetScan database, <http://www.targetscan.org/>
  - [18] miRWalk database, <http://www.unm.uni-heidelberg.de/apps/zmf/mirwalk/>
  - [19] MicroCosm Targets Version 5, [www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5](http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5)
  - [20] J. Yu *et al.*, Biochem. Biophys. Res. Commun. **349**, 59 (2006); M. Lu *et al.*, PLoS ONE **3**, e3420 (2008); G. Boross, K. Orosz and I. J. Farkas, Bioinformatics **25**, 1063 (2009).
  - [21] I. Ivanovska, M. A. Cleary, Cell Cycle **7**, 3137 (2008).
  - [22] S. Mookherjee *et al.*, Online J Bioinform. **10**, 280 (2009).
  - [23] J. Xu *et al.*, Nucleic Acids Res. **39**, 825 (2011).
  - [24] M. Basu, N. P. Bhattacharyya and P. K. Mohanty, J. Phys.: Conf. Ser. **297** 012002 (2011).
  - [25] M. Basu, N. P. Bhattacharyya and P. K. Mohanty, *arXiv:1307.1382*.